Course Title: **Information Systems Design** Course Code: CCE4235 4th year

Date: 9.4.2016 (Second term) Allowed time: 1 hrs

Answer Model:

Question No. 1 **(10 marks)**

For each of the following, please circle the letter introducing the best answer- each one is worth one mark:

1. What is an example of a null hypothesis?
   a) that a newly created model does not provide better predictions than the currently existing model
   b) that a newly created model provides a prediction of a null sample mean
   c) that a newly created model provides a prediction of a null population mean
   d) that a newly created model provides a prediction that will be well fit to the null distribution
   Answer a
2. In which phase of the data analytics lifecycle do Data Scientists spend the most time in a project?
   a) Discovery
   b) Data Preparation
   c) Model Building
   d) Communicate Results
   Answer b
3. You have used k-means clustering to classify behavior of 100, 000 customers for a retail store. You decide to use household income, age, gender and yearly purchase amount as measures. You have chosen to use 8 clusters and notice that 2 clusters only have 3 customers assigned. What should you do?
   a) Increase the number of clusters
   b) Decrease the number of measures used
   c) Decrease the number of clusters
   d) Identify additional measures to add to the analysis
   Answer c
4. In which lifecycle stage are initial hypotheses formed?
   a) Discovery
   b) Model planning
   c) Model building
   d) Data preparation
   Answer a
5. Consider the example of an analysis for fraud detection on credit card usage. You will need to ensure higher risk transactions that may indicate fraudulent credit card activity are retained in your data for analysis, and not dropped as outliers during pre-processing. What will be your approach for loading data into the analytical sandbox for this analysis?
   a) ELT
   b) ETL
   c) EDW
   d) OLTP
   Answer a

6. Which key role for a successful analytic project can consult and advise the project team on the value of end results and how these will be used on a day-to-day basis?
   a) Project Manager
   b) Business User
   c) Data Scientist
   d) Business Intelligence Analyst

   Answer b

7. A disk drive manufacturer, has a defect rate of less than 2% with 98% confidence. A quality assurance team samples 1000 disk drives and finds 14 defective units. Which action should the team recommend?
   a) The manufacturing process should be inspected for problems.
   b) A larger sample size should be taken to determine if the plant is functioning properly
   c) A smaller sample size should be taken to determine if the plant is functioning properly
   d) The manufacturing process is functioning properly and no further action is required.

   Answer d

8. Which characteristic applies only to Business Intelligence as opposed to Data Science?
   a) Supports solving "what if" scenarios
   b) Uses large data sets
   c) Uses only structured data
   d) Uses predictive modeling techniques

   Answer c

9. When would you use a Wilcoxson Rank Sum test?
   a) When you cannot make an assumption about the distribution of the populations
   b) When the data can easily be sorted
   c) When the populations represent the sums of other values
   d) When the data cannot easily be sorted

   Answer a

10. Which activity might be performed in the Operationalize phase of the Data Analytics Lifecycle?
    a) Try different analytical techniques
    b) Try different variables
    c) Transform existing variables
    d) Run a pilot

    Answer d

Question No. 2 **(15 marks)**

1. **What are the characteristics of Big Data?** **(2 marks)**

The characteristics of Big Data are:

1) **Volume (size)**
2) **Velocity (rapidly streaming)**
3) **Variety (many forms)**
4) **Veracity (Uncertainty of data)**
5) **Value of data (well and good for access or useless data)**

2. **Use the k-means algorithm and Euclidean distance to cluster the following eight examples into 3 clusters: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). The Euclidean distance is estimated using the**
٣ **equation:**

$$d\left((x_1, y_1), (x_2, y_2) \right) = \text{sqrt}\left((x_2-x_1)^2+(y_2-y_1)^2\right).$$

Suppose that the initial centers of each cluster are A1, A4 and A7. Run the k-means algorithm for 1 epoch (iteration) only. At the end of this epoch show:

a) The new clusters (i.e. the examples belonging to each cluster)

b) The centres of the new clusters

c) Draw a 10 by 10 space with all the 8 points and show the clusters after the first epoch and the new centroids.

d) How many more iterations are needed to converge? Draw the result for each epoch. **(4 marks)**

Solution:

a)

d(a,b) denotes the Eucledian distance between a and b. It is obtained directly from the distance matrix or calculated as follows: $d(a,b)=\text{sqrt}((x_b-x_a)^2+(y_b-y_a)^2)$

seed1=A1=(2,10), seed2=A4=(5,8), seed3=A7=(1,2)

epoch1 – start:

A1:

d(A1, seed1)=0 as A1 is seed1

d(A1, seed2)= $\sqrt{13}$ >0

d(A1, seed3)= $\sqrt{65}$ >0

➔A1 ∈ cluster1

A2:

d(A2,seed1)= $\sqrt{25}$ = 5

d(A2, seed2)= $\sqrt{18}$ = 4.24

d(A2, seed3)= $\sqrt{10}$ = 3.16 ← smaller

➔ A2 ∈ cluster3

A3:

d(A3, seed1)= $\sqrt{36}$ = 6

d(A3, seed2)= $\sqrt{25}$ = 5 ← smaller

d(A3, seed3)= $\sqrt{53}$ = 7.28

➔ A3 ∈ cluster2

A4:

d(A4, seed1)= $\sqrt{13}$

d(A4, seed2)=0 as A4 is seed2

d(A4, seed3)= $\sqrt{52}$ >0

➔ A4 ∈ cluster2

A5:

d(A5, seed1)= $\sqrt{50}$ = 7.07

d(A5, seed2)= $\sqrt{13}$ = 3.60 ← smaller

d(A5, seed3)= $\sqrt{45}$ = 6.70

➔ A5 ∈ cluster2

A6:

d(A6, seed1)= $\sqrt{52}$ = 7.21

d(A6, seed2)= $\sqrt{17}$ = 4.12 ← smaller

d(A6, seed3)= $\sqrt{29}$ = 5.38

➔ A6 ∈ cluster2

A7:

d(A7, seed1)= $\sqrt{65}$ >0

d(A7, seed2)= $\sqrt{52}$ >0

d(A7, seed3)=0 as A7 is seed3

➔ A7 ∈ cluster3

A8:

d(A8, seed1)= $\sqrt{5}$

d(A8, seed2)= $\sqrt{2}$ ← smaller

d(A8, seed3)= $\sqrt{58}$

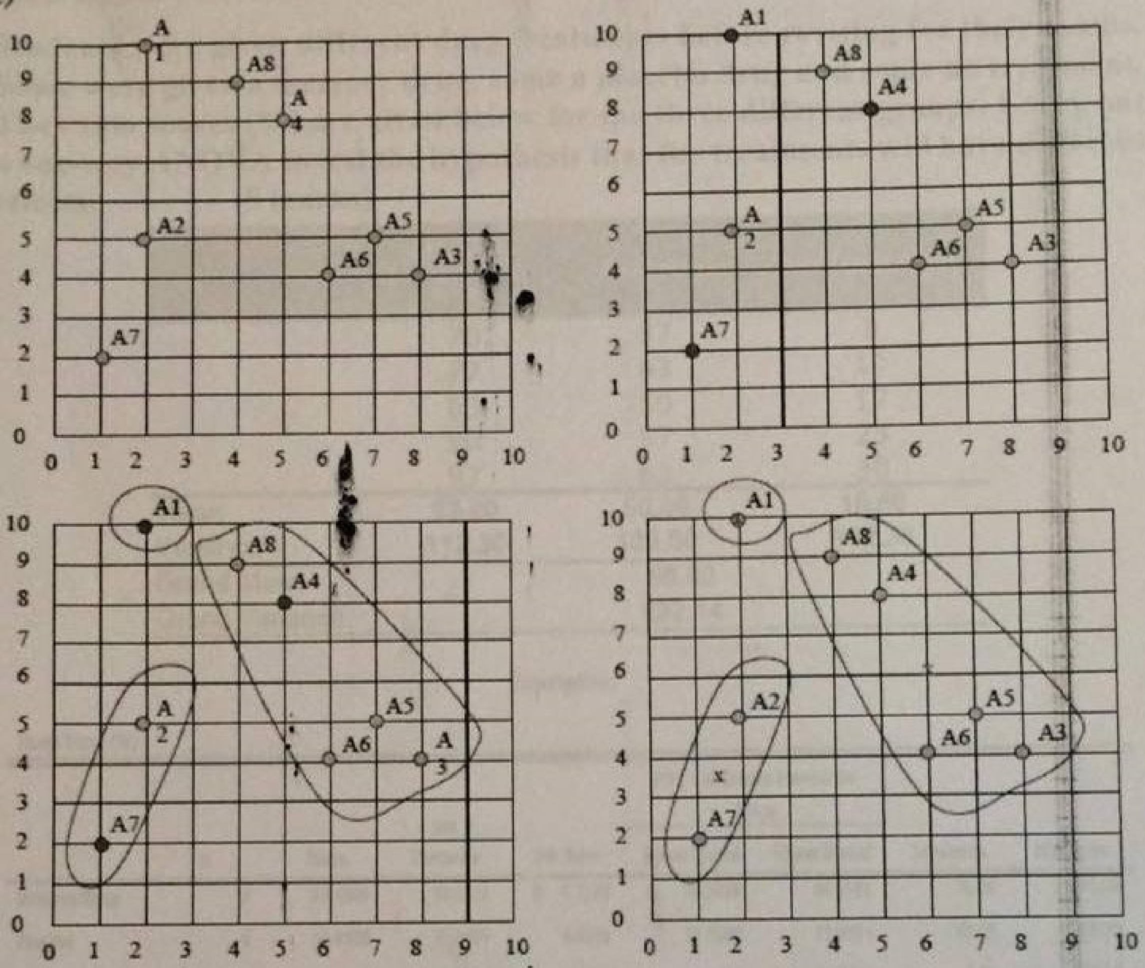
➔ A8 ∈ cluster2

end of epoch1

new clusters: 1: {A1}, 2: {A3, A4, A5, A6, A8}, 3: {A2, A7}

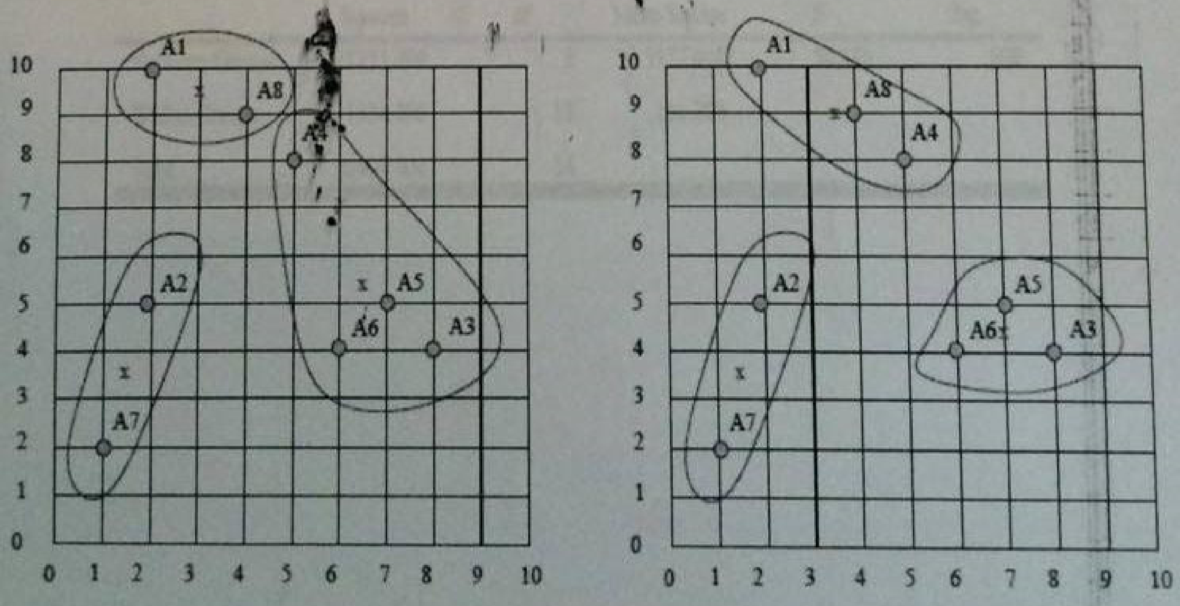
b) centers of the new clusters:

C1= (2, 10), C2= ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6), C3= ((2+1)/2, (5+2)/2) = (1.5, 3.5) c)



d)

We would need two more epochs. After the 2^{nd} epoch the results would be:
1: {A1, A8}, 2: {A3, A4, A5, A6}, 3: {A2, A7}
with centers C1=(3, 9.5), C2=(6.5, 5.25) and C3=(1.5, 3.5).
After the 3^{rd} epoch, the results would be:
1: {A1, A4, A8}, 2: {A3, A5, A6}, 3: {A2, A7}
with centers C1=(3.66, 9), C2=(7, 4.33) and C3=(1.5, 3.5).



3. **Students were given different drug treatments before revising for their exams. Some were given a memory drug, some a placebo drug and some no treatment. The exam scores (%) are given below for the three different groups. Carry out a one-way ANOVA to test the hypothesis that the treatments will have different effects.** (5 marks)

	Memory Drug	Placebo	No Treatment
	70	37	3
	77	43	10
	83	50	17
	90	57	23
	97	63	30
Mean	83.40	50.00	16.60
Variance	112.30	109.00	112.30
Grand Mean		50.00	
Grand Variance		892.14	

Descriptives

Exam Score (%)

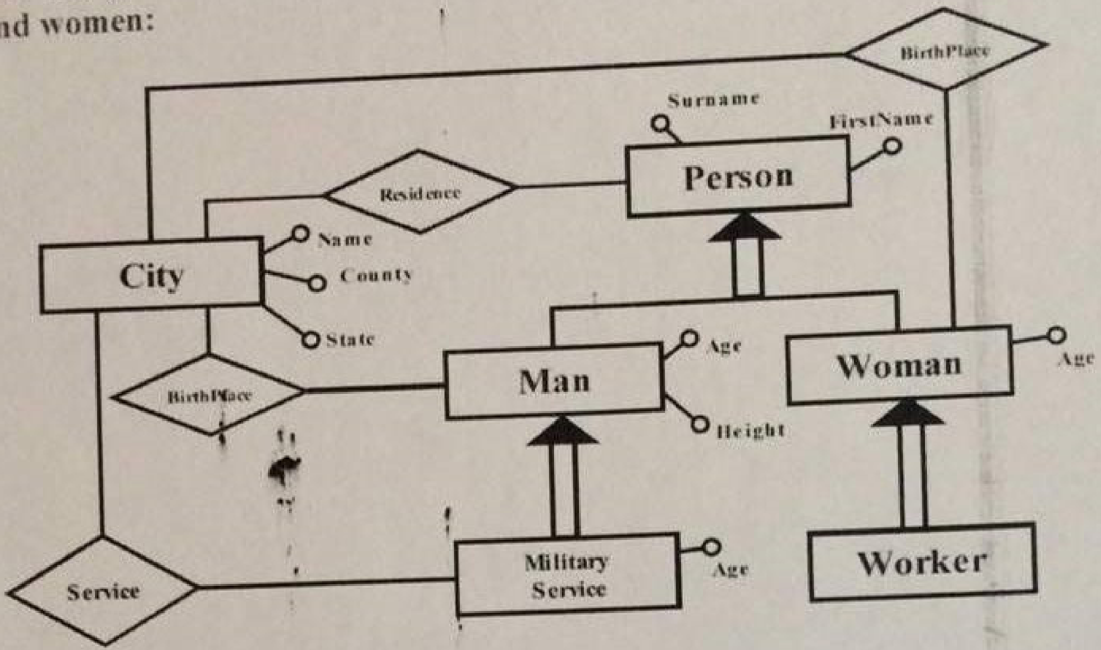
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Memory Drug	5	83.4000	10.5972	4.7392	70.2419	96.5581	70.00	97.00
Placebo	5	50.0000	10.4403	4.6690	37.0366	62.9634	37.00	63.00
No Treatment	5	16.6000	10.5972	4.7392	3.4419	29.7581	3.00	30.00
Total	15	50.0000	29.8688	7.7121	33.4592	66.5408	3.00	97.00

ANOVA

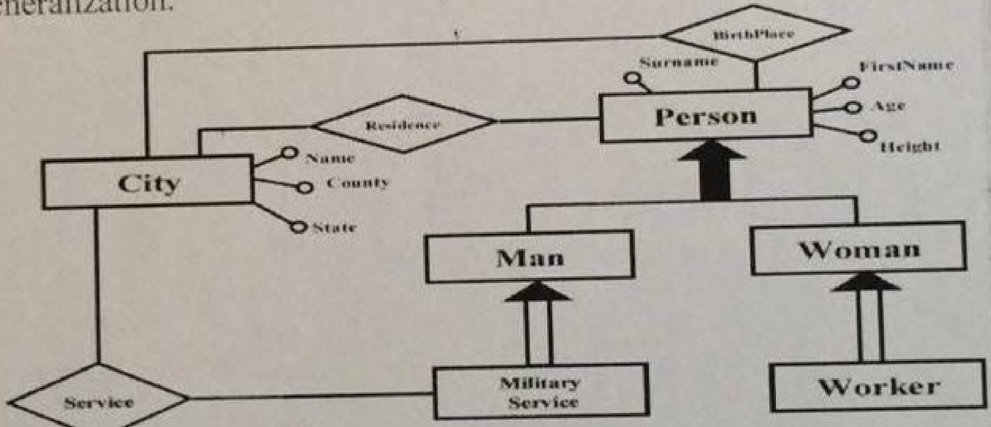
Exam Score (%)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	11155.600	2	5577.800	50.160	.000
Within Groups	1334.400	12	111.200		
Total	12490.000	14			

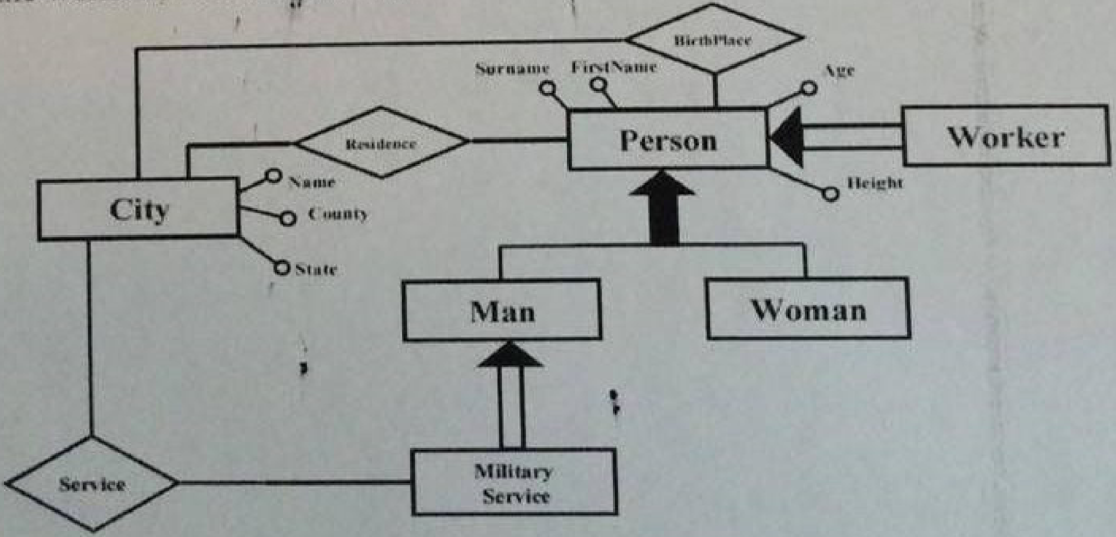
4. Consider the given E-R schema: the schema represents various properties of men and women: (4 marks)



a) Correct the schema, taking into account the fundamental properties of the generalization.



b) The schema represents only the female workers; modify the schema to represent all the workers, men and women.



Best wishes

Dr. Sherin El Gokhy